CWI
Centrum Wiskunde & Informatica

UAM
UNIVERSIDAD AUTONOMA
DE MADRID

Better Contextual Suggestions by Applying Domain Knowledge

Thaer Samar, Alejandro Bellogin,
Arjen P. de Vries

NWO

WebART
Web Archive Retrieval Tools

Contextual Suggestions

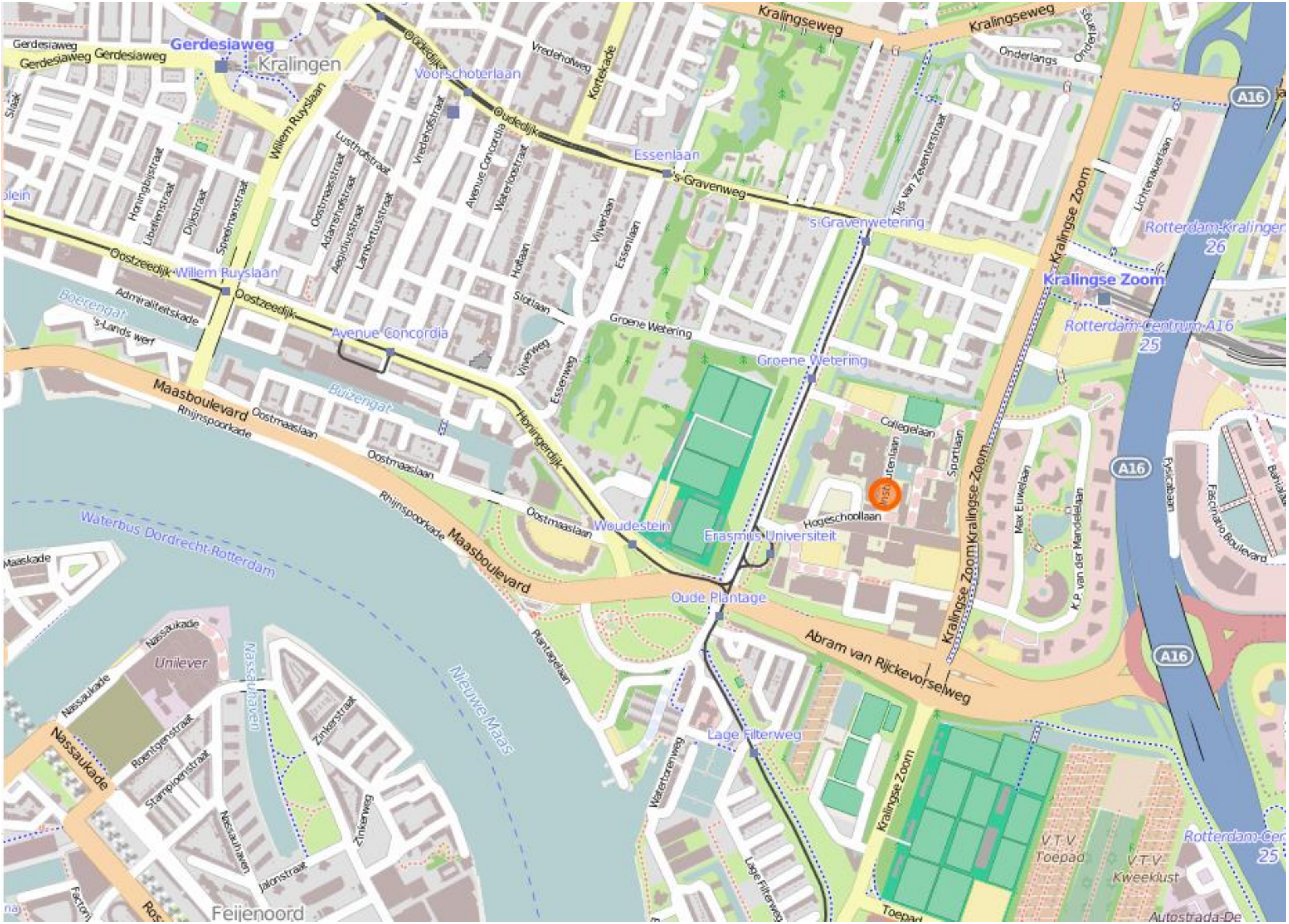
- Given a user profile and a context, make suggestions
    - *AKA* Context-aware Recommendation, zero-query Information Retrieval, …

“Entertain me”

- Recommend “things to do”, where
  - User profile consists of opinions about attractions
  - Context consists of a specific geo-location

My Profile

My Context



Gerdesiaweg
Kralingen
Kralingseweg
Onderlangs
Voorschoterlaan
Oudedijk
Vredeholweg
Kortekade
Willem Ruyslaan
Lusthofstraat
Essenlaan
's-Gravenweg
's-Gravenwetering
Tijs van Zeventerstraat
Kralingse Zoom
Lichtenauerlaan
Rotterdam-Kralingen
26
Honingbijstraat
Libellenstraat
Dijkstraat
Speelmanstraat
Oostmaasstraat
Adamshofstraat
Aegidiusstraat
Lambertusstraat
Avenue Concordia
Waterloostraat
Vijverlaan
Hoflaan
Slotlaan
Groene Wetering
Oostzeedijk
Admiraliteitskade
Boerengat
's-Lands werf
Buizengat
Vijverweg
Essenweg
Honingerdijk
Maasboulevard
Rhijnspoorkade
Oostmaaslaan
Collegelaan
Sportlaan
Burg. Oudlaan
Hogeschoollaan
Max Euwelaan
K.P. van der Mandelelaan
Fascinatio Boulevard
Fysicalaan
A16
Rotterdam-Centrum A16
25
Woudestein
Erasmus Universiteit
Oude Plantage
Abram van Rijckevorselweg
Waterbus Dordrecht-Rotterdam
Maaskade
Nassaukade
Unilever
Nassauhaven
Nieuwe Maas
Plantageplaan
Roentgenstraat
Stamplorenstraat
Zinkerstraat
Zinkerweg
Jalonstraat
Watertorenweg
Lage Filterweg
Toepad
V.T.V. Toepad
V.T.V. Kweekkust
Autostrada-De
Feijenoord

My Suggestions

WORM

Poortgebouw

TREC Contextual Suggestions (1/3)

- Given a user profile
  - 70 – 100 POIs represented by a title, description and URL (situated in Chicago / Santa Fe)
  - Rated on a scale 0 – 4

125, Adler Planetarium & Astronomy Museum, "Interactive exhibits & high-tech sky shows entertain stargazers -- lakefront views are a bonus.",
http://www.adlerplanetarium.org/
131,Lincoln Park Zoo,"Lincoln Park Zoo is a free 35-acre zoo located in Lincoln Park in Chicago, Illinois. The zoo was founded in 1868, making it one of the oldest zoos in the U.S. It is also one of a few free admission zoos in the United States.", http://www.lpzoo.org/

700, **125**, 4, 4
700, **131**, 0, 1

TREC Contextual Suggestions (2/3)

- … and a context
  - Corresponding to a metropolitan area in the USA, e.g., 109, Kalamazoo, MI

TREC Contextual Suggestions (3/3)

- Suggest Web pages / snippets
  - From the Open Web, or from ClueWeb

700, 109 ,1,"About KIA History Kalamazoo Institute of Arts KIA History","The Kalamazoo Institute of Arts is a nonprofit art museum and school. Since , the institute has offered art classes and free admission programming, including exhibitions, lectures, events, activities and a permanent collection. The KIAs mission is to cultivate the creation and appreciation of the visual arts for the communities",clueweb12-1811wb-14-09165

Approach

- For a given location, select candidate web pages from Clueweb
- Rank the candidates based on their cosine-similarity to the POIs in the user profile (separated in a positive and a negative profile)

Snippet Generation

- Generate POI title:
  - Extract `<title>` or `<header>` tags

- Generate *personalized* POI description:
  - Extract `<description>` tag
  - Break documents into sentences, ranked on their similarity with the user profile
  - Concatenate until 512 bytes reached

Candidate selection

- In 2013, the CWI Clueweb based run ranked far below all other (Open Web) runs
  - A few issues related to evaluation, see our ECIR 2014 short paper
  - But, also, the commercial Open Web search engines (Google, Bing or Yahoo!) return much better candidates for queries derived from the context than we did

Geo-Filtering

- Exact mention of given context
  - Format: {City, ST} e.g., Miami, FL

- Exclude documents that mention multiple contexts
  - E.g., a Wikipedia page about cities in Florida state

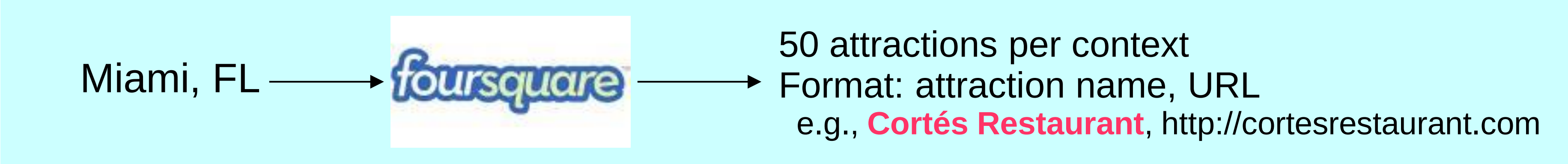
Domain Knowledge (1/2)

- Point-of-***Interest*** heuristic:
  - POIs will be represented on the major tourist information sites

    {`yelp`, `tripadvisor`, `wikitravel`, `zagat`, `xpedia`, `orbitz`, and `travel.yahoo`}
- Extract the Clueweb documents from these domains (TouristListFiltered)
  - E.g., http://www.zagat.com/miami
- Expand with the outlinks also contained in ClueWeb12 (TouristOutlinksFiltered)

Domain Knowledge (2/2)

- Use Foursquare API to identify the URLs of POIs for the given context



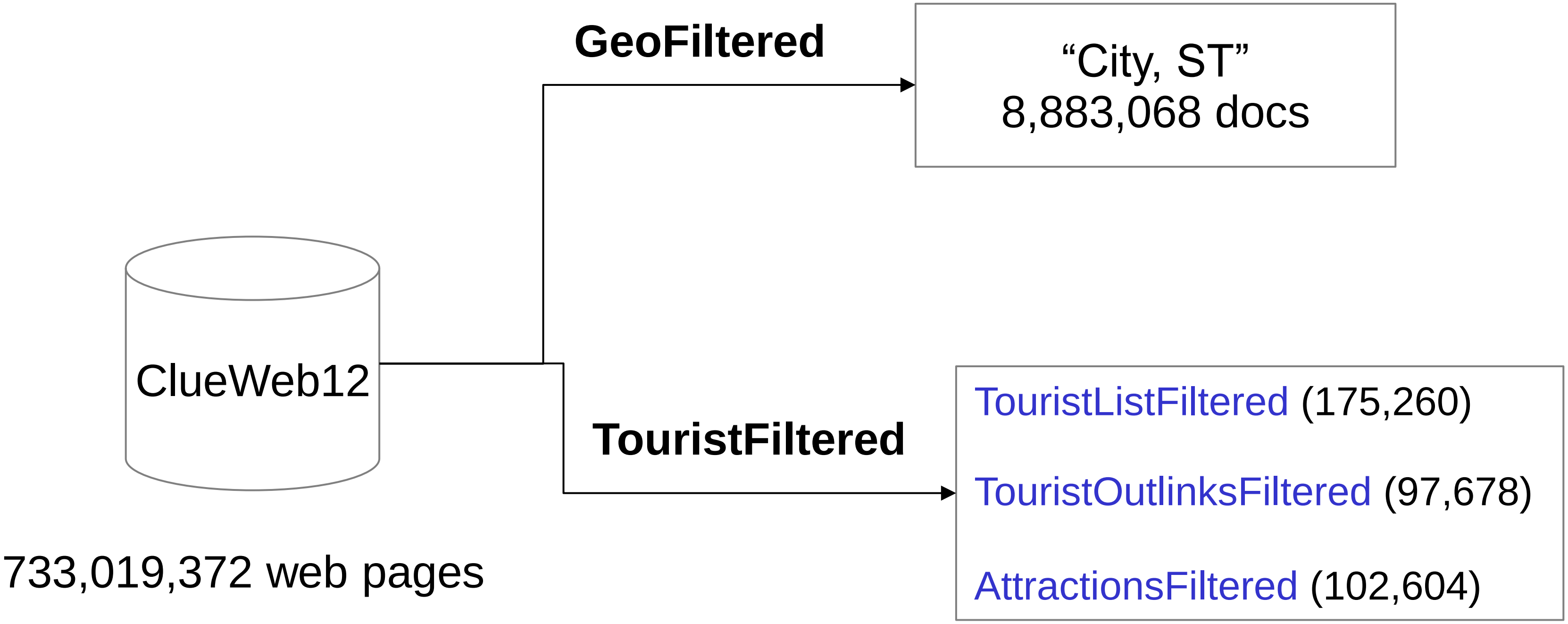
Miami, FL → foursquare →

50 attractions per context
Format: attraction name, URL
e.g., **Cortés Restaurant**, http://cortesrestaurant.com

  - If the POI has no corresponding URL, use Google API with a query using foursquare POI + context, i.e., “Cortés Restaurant Miami, FL”

- Extract any document from Clueweb whose host matches the (1,454 unique) hosts of the URLs identified (AttractionFiltered)

Candidate Selection



GeoFiltered

“City, ST”
8,883,068 docs

ClueWeb12

TouristFiltered

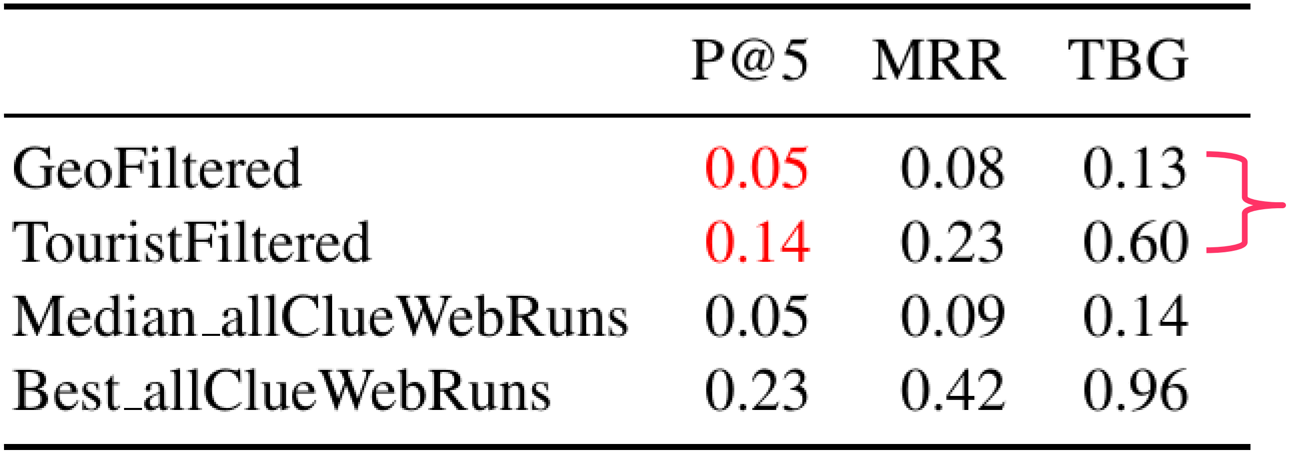
TouristListFiltered (175,260)

TouristOutlinksFiltered (97,678)

AttractionsFiltered (102,604)

733,019,372 web pages

Overall Results



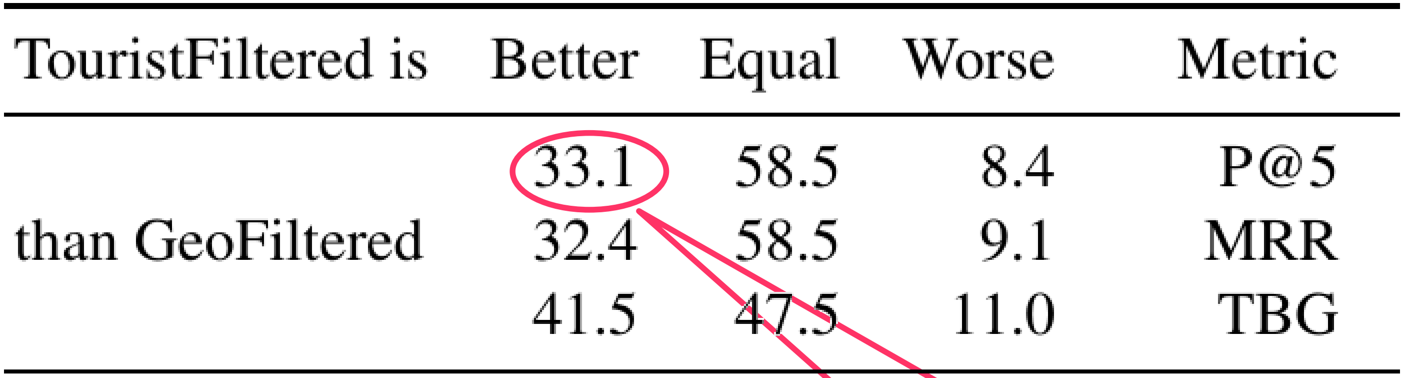
	P@5	MRR	TBG
GeoFiltered	0.05	0.08	0.13
TouristFiltered	0.14	0.23	0.60
Median_allClueWebRuns	0.05	0.09	0.14
Best_allClueWebRuns	0.23	0.42	0.96

TouristFiltered
>>
GeoFiltered

- Note:
  - P@5 and MRR consider three dimensions of relevance: geographical (geo), description (desc) and document (doc) relevance

TouristFiltered vs. GeoFiltered

% topics

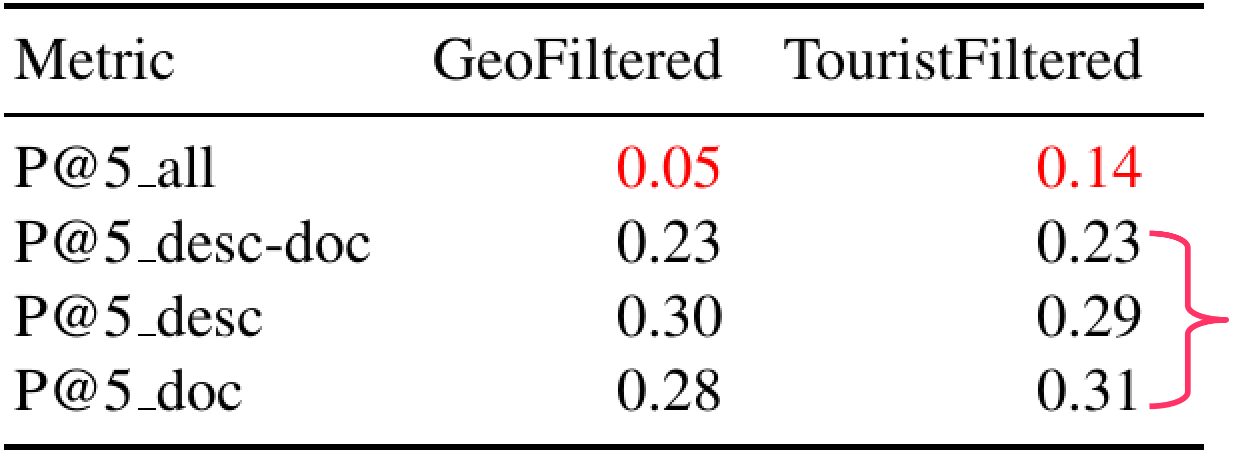


TouristFiltered is	Better	Equal	Worse	Metric
than GeoFiltered	33.1	58.5	8.4	P@5
	32.4	58.5	9.1	MRR
	41.5	47.5	11.0	TBG

(I.e., TouristFiltered suggests better POIs for 33.1% of the judged topics)

Decompose metrics

- Ignoring geo-relevance:

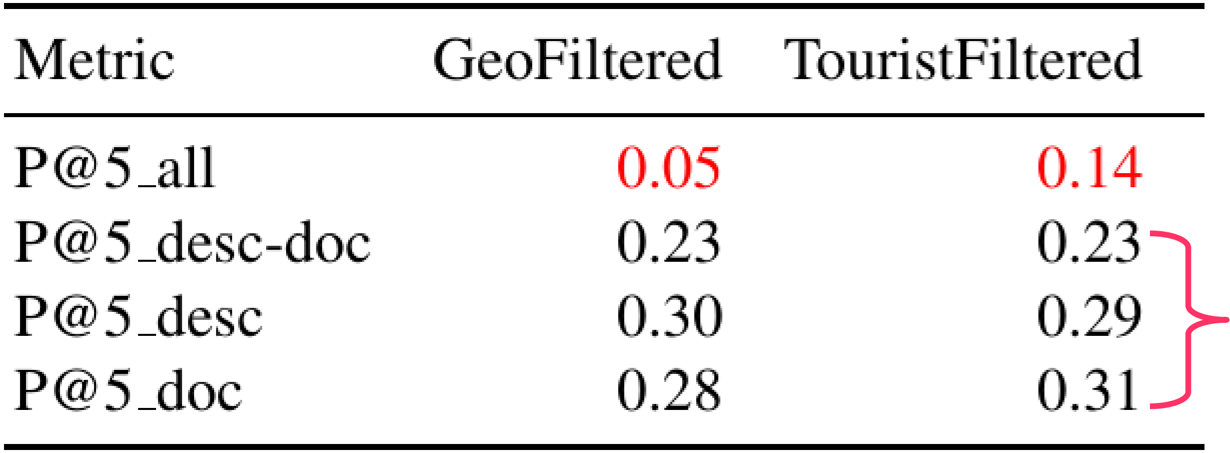


Metric	GeoFiltered	TouristFiltered
P@5_all	0.05	0.14
P@5_desc-doc	0.23	0.23
P@5_desc	0.30	0.29
P@5_doc	0.28	0.31

GeoFiltered ~ TouristFiltered

Decompose metrics

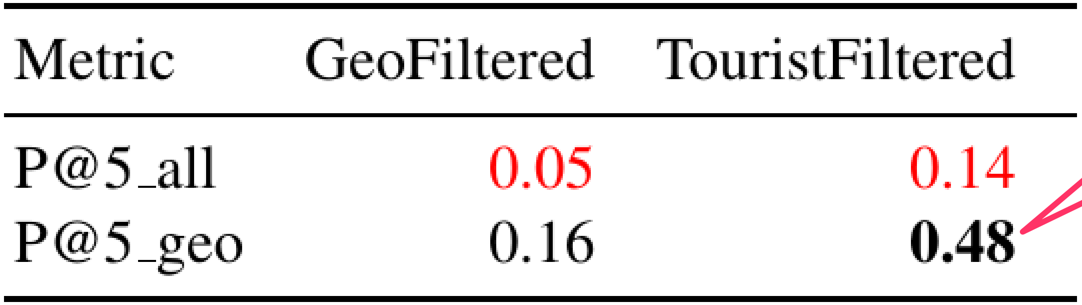
- Ignore geo-relevance:



Metric	GeoFiltered	TouristFiltered
P@5_all	0.05	0.14
P@5_desc-doc	0.23	0.23
P@5_desc	0.30	0.29
P@5_doc	0.28	0.31

The two runs have almost similar performance in the desc and doc dimensions

- Geo-relevance only:

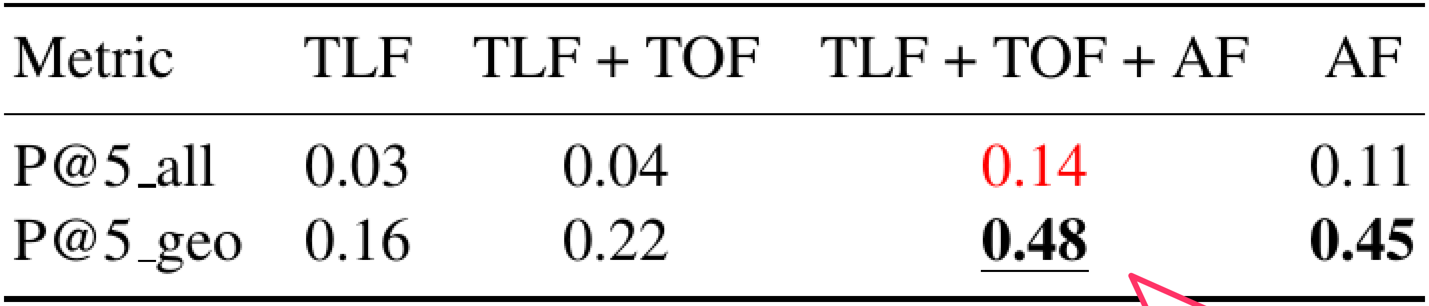


Metric	GeoFiltered	TouristFiltered
P@5_all	0.05	0.14
P@5_geo	0.16	**0.48**

TouristFiltered is more geographically appropriate

Type of domain knowledge

- TouristFiltered consists of three parts:
  - TouristListFiltered (TLF)
  - TouristOutlinksFiltered (TOF)
  - AttractionFiltered (AF)



Metric	TLF	TLF + TOF	TLF + TOF + AF	AF
P@5_all	0.03	0.04	0.14	0.11
P@5_geo	0.16	0.22	**<u>0.48</u>**	**0.45**

Foursquare gives the most significant improvement in performance

Conclusions

- Domain knowledge about sites that are more likely to offer attractions lead to better suggestions
- The best results were obtained when identifying attractions through specialized services such as Foursquare

Next Steps

- Improve our recommendation algorithm
  - E.g., weighted candidate selection
- Understand the remaining difference with Open Web based results
  - Our Clueweb results are **reproduceable** but not yet as good